Clinical Trials For Personalized Medicine: Designs and Some Statistical Challenges

Feifang Hu

Department of Statistics

University of Virginia

Email: fh6e@virginia.edu

Based on joint works with Yanqing Hu (UVa), Hongjian Zhu (Yale) and Hongyu Zhao (Yale)

Oct 25, 2011, IMS, NUS, Singapore

Outline

- Personalized Medicine
- Mathematical Framework
- Covariate-adaptive designs
- New covariate-adaptive designs for balance
- New CARA design for detecting interactions.
- Statistical inference and further research topics

1 Personalized medicine

From Wikipedia:
Personalized medicine is a medical model emphasizing the **systematic use of information about an individual patient** to select or optimize that patient's preventative and therapeutic care.

Personalized medicine can broadly be defined as **products and service that leverage the science of genomics and proteomics (directly and indirectly)** and capitalize on the trends toward wellness and consumerism to enable tailored approaches to prevention and care.

Over the past century, medical care has centered on standards of care based on epidemiological studies of large cohorts. However, large cohort studies do not take into account the genetic variability of individuals within a population. Personalized medicine (also call **Future medicine**) seeks to provide an objective basis for consideration of such individual differences.

Three main steps to develop personalized medicine:

- Identify important biomarkers that could be related with certain diseases: Bio-informatics, genomics, proteomics, and metabolomics, etc.
- Well designed clinical studies to confirm the significance of biomarkers to certain diseases and treatments, then approved by FDA.
- Implement to healthcare.

There are several stakeholders:

- The industry:
  - Pharmaceutical industry;
  - Diagnostics industry;
  - Insurers.
- Government agencies: FDA in USA.
- Both Physicians and Patients need to be educated.

In the past decades, fields of translational research (genomics, proteomics, and metabolomics) study the contribution of genes, proteins, and metabolic pathways to human physiology and variations of these pathways that can lead disease susceptibility.

Here are some recent examples:

- Khan, Fotheringham, Wood, *et al* (2010) reported a biomarker (HR23B) is highlighted as the main biomarker that could be used to determine CTCL (cutaneous T-cell lymphoma) cells' sensitivity to the drug SAHA (suberoylanilide hydroxamic acid). In their study, the researchers demonstrated how HR23B could be implemented as a biomarker in a clinically relevant setting. They showed that the presence of HR23B in biopsies from patients with CTCL predicted who would respond to the treatment 71.7% of the time.

- Li, Sheu, Ye, *et al* (2010) found that the top two SNPs (rs2352028 and rs235209) were connected with lung cancer in never smokers through their regulation of GPC5 expression.
- A recent study by Ashley, Butte, Matthew, Wheeler, *et al* (2010) indicated that rare variants in three genes that are clinically associate with sudden death- THEM43, DSP and MYBPC3. The study showed that genome sequenced data can be used to predict risk of diseases like myocardial infarction, type 2 diabetes and some cancers and response to treatments.

- In a study of McIlroy, McCartan, Early, *et al* (2010), Novel biomarkers (Nuclear HOXC11 and S100β) were found to predict poor disease-free survival in breast cancer patients, and these proteins could be detected in the blood.
- Genetic clues that may aid in the development of personalized medications for alcohol addiction as studied by Ramchandani, Umhau, Pavon, *et al* (2010). They found that mice with the 118G variant demonstrated a fourfold higher peak dopamine response.

- Lipkin, Chao, Moreno, *et al* (2010) found that HMGCR variant may help identify patients who are likely to benefit from Statins than others- for both cholesterol lowering and colorectal cancer prevention.

Identifying genes that seems to be linked with a disease in only **the first step** of developing personalized medicine. **New approaches** to the drug-development paradigm are needed, especially new designs for clinical trials so that genetics and other biomarkers can be incorporated to assist in patient and treatment selection.

On June 11, 2011, *The Economist* published a paper:

"**If personalized medicine is to achieve its full potential, it should be used earlier on in clinical trials.**"

Examples are discussed there.

What statisticians can do?:

- Formulate the procedure statistically (mathematically).
- Find optimal or efficient solutions.
- New Designs of clinical trials.
- New statistical inference tools.
- etc.

I will focus on the new designs of clinical trials in this talk.

Complexness of data structure:

- Many covariates:
  - biomarkers;
  - investigation sides;
  - other covariates (male or female; smoker or nonsmoker, etc);
- sequentially dependent.
- missing data.
- small sample size.
- multi-treatments.
- others.

2 Mathematical framework

Consider a clinical trial of n patients, each of whom is to randomly receive one of 2 treatments (can be generalized to multi-treatments). A *randomization sequence* is a matrix $\boldsymbol{T} = (T_1, ..., T_n)'$, where $T_i = 1$ if patient i is in treatment 1 and $T_i = 0$ if patient i is assigned in treatment 2 for $i = 1, ..., n$ $N_1(n) = \sum_{i=1}^{n} T_i$ is the total number of patients in treatment 1. $N_2(n) = n - N_1(n)$.

Let $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)'$, where $\boldsymbol{X}_i = (X_{i1}, X_{i2})$, be a matrix of response variables, where $\boldsymbol{X}_i$ represents the sequence of responses that would be observed if each treatment were assigned to the i-th patient independently. However, only one element of $\boldsymbol{X}_i$ will be observable. $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ are their corresponding covariate vectors (biomarkers, etc.). Throughout the talk, we will consider probability models for $\boldsymbol{X}_i$ conditional on $\boldsymbol{T}_i$ and $\boldsymbol{Z}_i$.

The data structure: $\{\boldsymbol{Z}_i, \boldsymbol{T}_i, \boldsymbol{X}_i\}$, $i = 1, ..., n$.

- Let $\mathcal{T}_n = \sigma\{T_1, ..., T_n\}$ be the sigma-algebra generated by the first n treatment assignments,
- let $\mathcal{X}_n = \sigma\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ be the sigma-algebra generated by the first n responses,
- let $\mathcal{Z}_n = \sigma\{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n\}$ be the sigma-algebra generated by the first n covariate vectors.
- Let $\mathcal{F}_n = \mathcal{T}_n \otimes \mathcal{X}_n \otimes \mathcal{Z}_{n+1}$.

A *randomization procedure* is defined by

$$\phi_n = E(\boldsymbol{T}_n | \mathcal{F}_{n-1}),$$

where ϕ_{n+1} is $\mathcal{F}_n$-measurable. We can describe ϕ_n as the conditional probability of assigning treatments 1 to the n-th patient, conditional on the previous $n-1$ assignments, responses, and covariate vectors, and the current patient's covariate vector.

We can describe five types of randomization procedures. We have

- *complete randomization* if

$$\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n);$$

- *restricted randomization* if

$$\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n|\mathcal{T}_{n-1});$$

- *response-adaptive design (randomization)* if

$$\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n|\mathcal{T}_{n-1}, \mathcal{X}_{n-1});$$

- *covariate-adaptive randomization (design)* if

$$\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n|\mathcal{T}_{n-1}, \mathcal{Z}_n);$$

- *covariate-adjusted response-adaptive (CARA) randomization (design)* if

$$\phi_n = E(\boldsymbol{T}_n|\mathcal{F}_{n-1}) = E(\boldsymbol{T}_n|\mathcal{T}_{n-1}, \mathcal{X}_{n-1}, \mathcal{Z}_n).$$

3 Covariate-adaptive design

Clinical trialists are often concerned that treatment arms will be unbalanced with respect to key covariates of interest. To prevent this, covariate-adaptive randomization is often employed. Over 50000 covariate-adaptive clinical trials had been reported from 1988-2008 (Taves, 2010).

Two popular procedures:

- *Stratified Block Randomization*: Use permuted block designs within each stratum. (about 95% reported trials).
- *Pocock-Simon procedure* (1975, Biometrics) (based on the biased coin idea from Efron (1971)). (about 5% trials, but increasing trend).

Stratified Block Randomization: Use permuted block designs within each stratum.

Permuted Block Design: permutation of m A's and m B's.

- e.g.: block size $2m = 4$, permutation of (AABB) or (BAAB);
For 10 patients: —AABB—BAAB—BB

- Advantage:
  - Easy to understand and implement.
  - Good large sample properties (almost prefect balance).
  - Balance within stratum.
- Disadvantage:
  - Only consider balance within stratum.
  - Does not work for cases with many strata (many covariates or many levels).

Pocock-Simon procedure: Let $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ be the covariate vector of patients $1, ..., n$. Assume that there are S covariates of interest (continuous or otherwise) and they are divided into $n_s, s = 1, ..., S$, different levels.

$N_{sik}(n), s = 1, ..., S, i = 1, ..., n_s, k = 1, 2$ to be the number of patients in the i-th level of the s-th covariate on treatment k.

Let patient $n + 1$ have covariate vector $\boldsymbol{Z}_{n+1} = (r_1, ..., r_S)$.

Let $D_s(n) = N_{sr_s1}(n) - N_{sr_s2}(n)$, which is the difference between the numbers of patients on treatments 1 and 2 for members of level r_s of covariate s.

Let $w_1, ..., w_S$ be a set of weights and take the weighted aggregate $D(n) = \sum_{s=1}^{S} w_s D_s(n)$. Establish a probability $\pi \in (1/2, 1]$. Then the procedure allocates to treatment 1 according to

$$\begin{aligned} \phi_{i1} = E(T_{i1}|\mathcal{T}_{i-1}, \mathcal{Z}_i) &= 1/2, \quad \text{if} \quad D(i-1) = 0, \\ &= \pi, \quad \text{if} \quad D(i-1) < 0, \\ &= 1-\pi, \quad \text{if} \quad D(i-1) > 0. \end{aligned}$$

- Advantage:
  - Balance across covariates (marginal balance).
  - Overall treatment balance with many covariates.
- Disadvantage:
  - Unknown theoretical properties (not well studied, Rosenberger and Sverdlov, 2009).
  - usually not well balanced within stratum.

We need new covariate-adaptive designs that provide balance (within stratum, marginal and overall) under different situations (sample size 200, 500 or 1000):

- 10 covariates, each with 2 levels: total $2^{10} = 1024$ strata.
- 2 covariates: a biomarker with 2 levels and 100 investigation sides: total 200 strata.

4 New Covariate-Adaptive Designs for Balance

Consider two covariates: covariate 1 with I levels and covariate 2 with J levels, For patient $n+1$ (with i (covariate 1) and j (covariate 2)) $n = 0, 1, 2, ...$ First we define the following values:

- If patient $n+1$ is assigned to treatment 1, let
  - *Within Stratum*: $D_{ij}^{(1)}(n+1) = N_{ij,1}(n+1) - N_{ij,2}(n+1)$, where $N_{ij,1}(n+1)$ and $N_{ij,2}(n+1)$ are the number of patients assigned to treatment 1 and 2 respectively in strata ij of the first $n+1$ patients.
  - *Marginal 1:* $D_{i\cdot}^{(1)}(n+1) = N_{i\cdot,1}(n+1) - N_{i\cdot,2}(n+1)$, where $N_{i\cdot,1}(n+1)$ and $N_{i\cdot,2}(n+1)$ are the number of patients assigned to treatment 1 and 2 respectively in (covariate 1=i) of the first $n+1$ patients.

- *Marginal 2:* $D_{\cdot j}^{(1)}(n+1) = N_{\cdot j,1}(n+1) - N_{\cdot j,2}(n+1)$, where $N_{\cdot j,1}(n+1)$ and $N_{\cdot j,2}(n+1)$ are the number of patients assigned to treatment 1 and 2 respectively in (covariate 2=j) of the first $n+1$ patients.
- *Overall:* $D_{n,overall} = N_{n,1} - N_{n,2}$ be the overall difference of patient numbers in group 1 and 2 among the first n.
- Define $A_{ij}^{(1)}(n+1) = (D_{ij}^{(1)}(n+1))^2$, $A_{i\cdot}^{(1)}(n+1) = (D_{i\cdot}^{(1)}(n+1))^2$, $A_{\cdot j}^{(1)}(n+1) = (D_{\cdot j}^{(1)}(n+1))^2$ and $A_{\cdot\cdot}^{(1)} = (D_{n,overall})^2$.
- The score of imbalance is $B_{ij}^{(1)}(n+1) = w_1 A_{ij}^{(1)}(n+1) + w_2 A_{i\cdot}^{(1)}(n+1) + w_3 A_{\cdot j}^{(1)}(n+1) + w_4 A_{i\cdot}^{(1)}(n+1)$ for some weights $w_1, w_2, w_3, w_4 \geq 0$.

- If patient $n+1$ is assigned to treatment 2, $B_{ij}^{(1)}(n+1)$ is calculated similarly.

Then the proposed procedure allocates to treatment 1 according to

$$\begin{aligned}
\phi_{n+1,1} &= 1/2, \quad \text{if} \quad B_{ij}^{(1)}(n+1) = B_{ij}^{(2)}(n+1), \\
&= \pi, \quad \text{if} \quad B_{ij}^{(1)}(n+1) < B_{ij}^{(2)}(n+1), \\
&= 1-\pi, \quad \text{if} \quad B_{ij}^{(1)}(n+1) > B_{ij}^{(2)}(n+1).
\end{aligned}$$

Where $\pi > 0.5$ ($\pi \in (0.75, 0.95)$ is recommended).

Remarks:

- When weight $w_1 = 0,\ w_4 = 0$, the new design becomes Pocock and Simon's procedure.
- When $w_2 = w_3 = w_4 = 0$, the new design is similar to Stratified Block Randomization.
- With $w_1, w_2, w_3 > 0$, we can balance both within each strata and cross covariates.

Theorem: Under certain conditions ($w_1 > 0$ and some others), $\mathbf{D}_n$ (imbalance matrix) is a positive recurrent Markov chain.

The proof is quite difficult because the correlated structure. This theorem ensures good balance for both within strata and cross factors (marginal).

Some numerical results:

Case 1: 10 covariates, each with 2 levels: total $2^{10} = 1024$ strata.

Table 1. Averaging imbalance under 100 simulations and $n = 500$

Dist of pts across strata			Counts & percentages		
# of pts	E(# prop)	Imb	strt(PB)	P-S	New
2	.07	0	50.2(.67)	38.2(.50)	55.1(.74)
		2	24.9(.33)	37.8(.50)	18.9(.26)
3	.01	1	12(1.00)	9.3(.77)	12.0(.96)
		3	0(0.00)	2.8(.23)	.5(.04)
(< 2)	.91				
	overall abs dif		12.8	.76	.90
	margnal abs dif		10.4	1.68	1.90

Table 2. Averaging imbalance under 100 simulations and $n = 1000$

Dist of pts across strata			Counts & percentages		
# of pts	E(# prop)	Imb	strt(PB)	P-S	New
2	.18	0	123.4(.67)	93.5(.51)	135.0(.74)
		2	60.0(.33)	91.0(.49)	48.0(.26)
3	.06	1	59.89(1.00)	45.1(.76)	57.1(.95)
		3	0(0.00)	14.4(.24)	2.8(.05)
(< 2)	.75				
(> 5)	.00				
overall abs dif			19.46	.62	1.1
margnal abs dif			14.29	1.61	2.1

Case 2: 2 covariates: a biomarker with 2 levels and 100 investigation sides: total 200 strata.

Table 3. Averaging imbalance under 1000 simulations and $n = 200$

Dist of pts across strata			Counts & percentages		
# of pts	E(# prop)	Imb	strt(PB)	P-S	New
2	.184	0	24.46(.66)	24.15(.65)	30.23(.82)
		2	12.37(.34)	12.74(.35)	6.46(.18)
3	.06	1	12.02(1.00)	11.18(.92)	12.05(.97)
		3	0(0.00)	1.02(0.08)	0.35(.03)
(< 2)	.735				
overall abs dif			9.39	1.14	1.53
margnal long abs dif			6.57	0.87	1.13
margnal short abs dif			1.00	0.86	0.81

Table 4. Averaging imbalance under 1000 simulations and $n = 500$

Dist of pts across strata			Counts & percentages		
# of pts	E(# prop)	Imb	strt(PB)	P-S	New
2	.257	0	34.24(.67)	31.46(.61)	41.51(.81)
		2	17.14(.33)	19.99(.39)	9.96(.19)
3	.214	1	42.84(1.00)	37.76(.88)	41.47(.97)
		3	0(0.00)	5.21(0.12)	1.19(.03)
(< 2)	.286				
overall abs dif			10.25	1.31	1.71
margnal long abs dif			7.23	0.96	1.28
margnal short abs dif			1.02	0.92	0.87

5 Optimal design for detecting interactions among treatments and biomarkers.

Covariate-adjusted response-adaptive (CARA) randomization (Zhang, Hu, Cheung and Chan, 2007, Annals; Rosenberger and Sverdlov, 2009, Statistical Science, etc.).

The main feature of CARA randomization is that patients are allocated on the basis of previous responses *and* the previous and current patient's known covariate profile.

Such procedures allow patients to be assigned to the treatment that is better for their individual covariate profile.

The goal of a conventional clinical trial is to determine if a new treatment is superior. When designing a clinical trial for personalized medicine, our goal is not limited to detect the treatment difference, but also to identify biomarkers that predict efficacy of treatments. Therefore, the interaction between the treatment and the biomarker becomes especially important.

Consider the linear model,

$$E(X_i) = \beta_0 + \beta_1 Z_{i1} +, \cdots, + \beta_K Z_{iK} + \beta_T T_i + \beta Z_{i1} T_i, \ i = 1, \ldots, n \tag{1}$$

where the X_i's are independent with error from normal distribution $N(0, \sigma_i^2)$,
$Z_{i1}, \ldots, Z_{iK}$ are covariates,
T_i is the treatment assignment taking values 1 or 0 for treatment 1 or 2 respectively
and $(\beta_0, \beta_1, \ldots, \beta_K, \beta_T, \beta)$ are the unknown parameters.
β is the interaction between treatment and covariate Z_1 (taking values 1 or 0 only).

In the model, we are interested in the effect of the interaction term, that is.

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0. \tag{2}$$

Our objective is to find the optimal allocation to maximize the power for the above hypothesis test and propose an effective design to target the optimal allocation.

New approach: The patients are divided into four groups based on treatment assignment and the value of covariate Z_1.

- Group (1) contains patients in treatment A with $Z_1 = 1$;
- group (2) contains patients in treatment A with $Z_1 = 0$;
- group (3) contains patients in treatment B with $Z_1 = 1$;
- group (4) contains patients in treatment B with $Z_1 = 0$.

Let $X^j, j = 1, 2, 3, 4$, be the response for four independent groups. Assume the variance of responses in the four groups to be σ_j^2, $j = 1, 2, 3, 4$, respectively. Let $X^j, j = 1, 2$ are the responses for treatment 1 and 2, $\mathbf{Z}_i^j = (Z_{i1}, \ldots, Z_{iK})^T$ are covariate vectors for the ith patient in group (j), n_j is the number of the patients in group (j), $j = 1, 2.$, and $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_K)^T$.

Denote n_j be the number of the patients in group (j), we have the following result:

Theorem: *Consider the linear model with specified variance as above and hypothesis (2), the optimal allocation for maximizing the power requires both of the following conditions:*

(A) $\sum_{i=1}^{n_1} \boldsymbol{Z}_i^1 \boldsymbol{\beta}/n_1 = \sum_{i=1}^{n_2} \boldsymbol{Z}_i^2 \boldsymbol{\beta}/n_2 = \sum_{i=1}^{n_3} \boldsymbol{Z}_i^3 \boldsymbol{\beta}/n_3 = \sum_{i=1}^{n_4} \boldsymbol{Z}_i^4 \boldsymbol{\beta}/n_4$

(B) $\frac{n_1}{n_1+n_3} = \frac{\sigma_1}{\sigma_1+\sigma_3}$ *and* $\frac{n_2}{n_2+n_4} = \frac{\sigma_2}{\sigma_2+\sigma_4}$

This can be viewed as a generalization of Neyman allocation.

CARA Randomization: Suppose n data points $(y_i, z_{i1}, \ldots, z_{iK}, T_i,\ i = 1, \ldots, n)$ have been observed. When the $(n+1)$th patient with covariates $(z_{(n+1)1}, \ldots, z_{(n+1)K})$ enters the trial,

(A) Obtain the estimator $\hat{\boldsymbol{\beta}}''_n$ of parameter $\boldsymbol{\beta}'' = (\beta_0, \beta_2, \ldots, \beta_K)$ and the estimator $\hat{\sigma}_i, i = 1, 2, 3, 4$ of standard deviations for the four groups by least squares.

(B) Count the number of patients in each of the four groups, i.e. (n_1, n_2, n_3, n_4).

(C) Assume the $(n+1)$th patient is assigned to treatment 1. Calculate the variance (VAR_1) of the four items $(\frac{\sum \boldsymbol{Z}_i^1 \hat{\boldsymbol{\beta}}_n''}{n_1}, \frac{\sum \boldsymbol{Z}_i^2 \hat{\boldsymbol{\beta}}_n''}{n_2}, \frac{\sum \boldsymbol{Z}_i^3 \hat{\boldsymbol{\beta}}_n''}{n_3}, \frac{\sum \boldsymbol{Z}_i^4 \hat{\boldsymbol{\beta}}_n''}{n_4})$.

(D) Suppose the $(n+1)$th patient is assigned to treatment 2. Calculate the corresponding variance (VAR_2) in the same way as in step (C). If $z_{(n+1)1} = 1$, go to step (E1), otherwise go to step (E2).

(E1) Calculate $D = w_1(VAR_1 - VAR_2) + w_2(\frac{n_1}{n_1+n_3} - \frac{\hat{\sigma}_1}{\hat{\sigma}_1+\hat{\sigma}_3})$, where $w_1, w_2 > 0$.

(E2) Calculate $D = w_1(VAR_1 - VAR_2) + w_2(\frac{n_2}{n_2+n_4} - \frac{\hat{\sigma}_2}{\hat{\sigma}_2+\hat{\sigma}_4})$.

(F) Assign the next patient to treatment 1 with the following probability

$$\psi = \begin{array}{cc} \pi & D < 0 \\ 0.5 & D = 0 \\ 1-\pi & D > 0 \end{array} , \tag{3}$$

where $\pi > 0.5$ (usually $\pi \in (0.75, 0.95)$.

Simulation: 1000 simulations with $n = 500$ data points from the model with three covariates; $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_T) = (1, 10, 5, 3, 8)$, $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 1, 2, 2)$. By simulation, we also found that our method could save over 10% of sample size.

Table 1: Table 5. Comparison of power

Randomization	β	power	$\hat{\beta}$ (s.e.)	$n_1/(n_1+n_3)$ (s.e.)
CR	0.5	0.698	0.491 (0.198)	0.500 (0.023)
NEW	0.5	0.748	0.502 (0.190)	0.334 (0.014)
CR	0.6	0.834	0.589 (0.198)	0.500 (0.022)
NEW	0.6	0.890	0.598 (0.180)	0.335 (0.015)

6 Statistical Inference and Some Further Problems

- Statistical Inference under covariate-adaptive randomization:
  - A theory for testing hypotheses (Shao, Yu and Zhang, 2010, Biometrika) for special cases.
  - Re-randomization tests (Rosenberger and Lachin, 2002, Jeon and Hu, 2010, etc.) for some simple situations.
  - Due to the complexness of the data structure, new methods?
  - Subgroup (certain gene type) statistical analysis.

- How to sequentially monitor covariate-adaptive randomized clinical trials? Zhu and Hu (2010, Annals) considered how to sequential monitor response-adaptive randomized clinical trials.
- Interim studies of covariate-adaptive randomized clinical trials
- Covariate-adjusted response-adaptive designs.
- multi-treatment.
- Many new problems.

Thank you!